ASSIGNMENT - 4

Assignment Report: Review Summarization using GPT-2

Code Structure:

1. Data Preprocessing: The initial section of the code involves reading the review dataset from a CSV file, preprocessing the data, and preparing it for training. This includes dropping unnecessary columns, handling missing values, truncating text to fit within the model's input size, and splitting the dataset into training and testing sets.

2. Model Initialization: The code initializes the GPT-2 model and tokenizer from the Hugging Face library. It also configures the device for training based on GPU availability.

3. Dataset Preparation: The code defines a custom dataset class to prepare the data for training. Each review is tokenized using the tokenizer and encoded into integer sequences. These sequences are then padded to ensure uniform input size.

4. Model Training: The training arguments are set, including the output directory, batch size, number of epochs, learning rate, and whether to use mixed precision. The Trainer class from the Transformers library is used to train the GPT-2 model on the prepared dataset.

5. Model Evaluation: After training, the model is evaluated using the test dataset. The code generates summaries for a subset of test reviews and calculates evaluation metrics such as ROUGE scores to assess the quality of the generated summaries.

Key Components:

- Tokenizer and Model: The GPT-2 tokenizer and model are initialized and configured for training and inference.
- Custom Dataset Class: A custom dataset class is defined to preprocess and prepare the data for training.
- Training Arguments: Various training arguments such as batch size, epochs, and learning rate are specified to train the model.
- Trainer: The Trainer class is used to facilitate model training with the specified training arguments and dataset.
- Evaluation: Evaluation metrics, particularly ROUGE scores, are calculated to measure the summarization quality compared to the ground truth summaries.

Results and Evaluation:
The code generates summaries for a subset of test reviews and calculates ROUGE scores as evaluation metrics. These scores provide insights into the performance of the model in generating accurate and informative summaries compared to the actual summaries provided in the dataset.

Sample Output

```
Article:  We love these organic corn chips. They have an excellent crunch big chips great for dipping. Try with  Salsa, Mild, Organic, 17.5 oz. or homemade salsa. We hav
Actual Summary:  Most Excellent Organic yellow corn chips. Great with homemade salsa.
Predicted Summary:   Delicious! Great taste and great price! Great for dipping. Great for dipping in salsa. Great for dipping in salsa.
```